Final Project Written Report

Julia Etzkorn, Izabelle Baysinger, Nathan Petersen

STAT 4511

April 29, 2025

**Introduction**

Predicting health insurance charges is notoriously complicated: accidents can happen, diseases can unexpectedly worsen, and treatments can take multiple rounds to be effective. With the seemingly random medical events that can occur, we wanted to see if it was possible to fit an accurate model to insurance charges. In this project, we test whether insurance charges can be predicted by a multivariate dataset to produce an effective linear model. Health insurance charges in the U.S. have been observed as they relate to age, sex, BMI, number of children, smoking habits, and region based on a dataset from kaggle.com. This dataset was originally derived as a synthetic dataset for model fitting based on US census data.

In this set, the mean age for policyholders is 39.21 years, with a minimum age of 18 years and a maximum age of 64 years. **Figure A** reveals a symmetric distribution with no outliers. The mean BMI for policyholders is 30.66, with a minimum of 15.96 and a maximum of 53.13. **Figure B** reveals a near normal distribution with the exception of a few high outliers. The distribution of policyholders’ sex is 50.523% male (coded 1) and 49.477% female (coded 0).

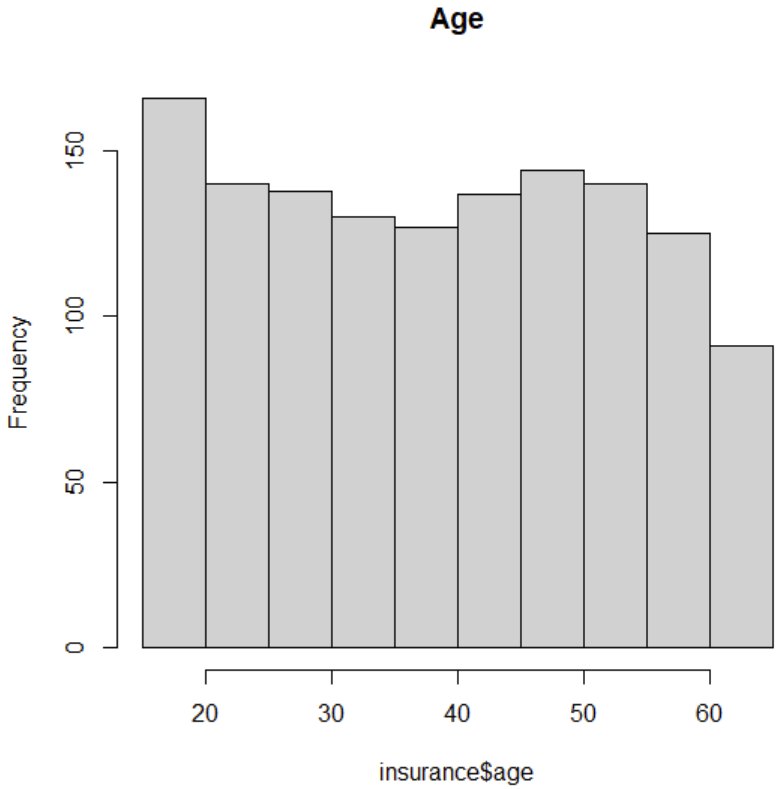
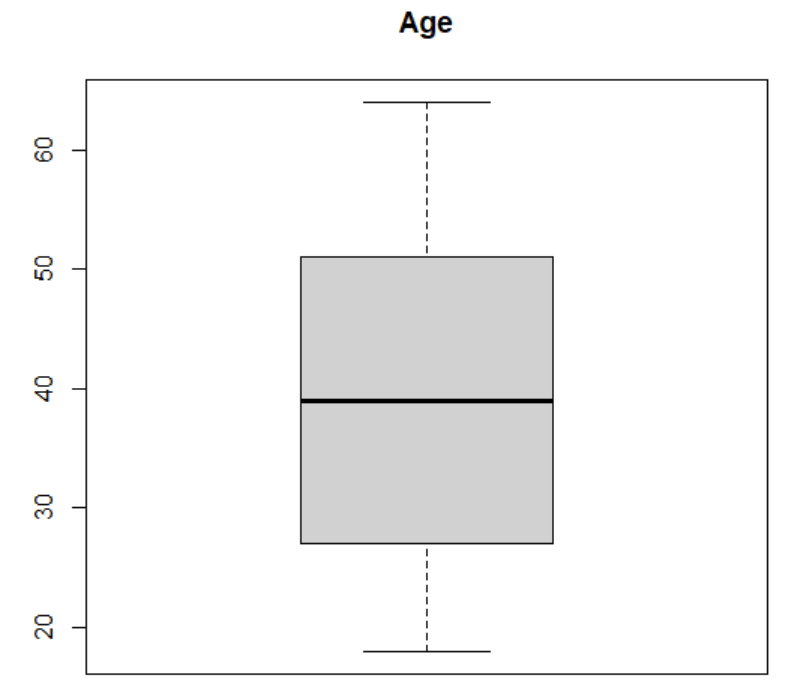


Figure A: a box plot and histogram of age. The histogram on the right reveals a symmetric distribution, while the box plot on the left confirms that the dataset contains no outliers

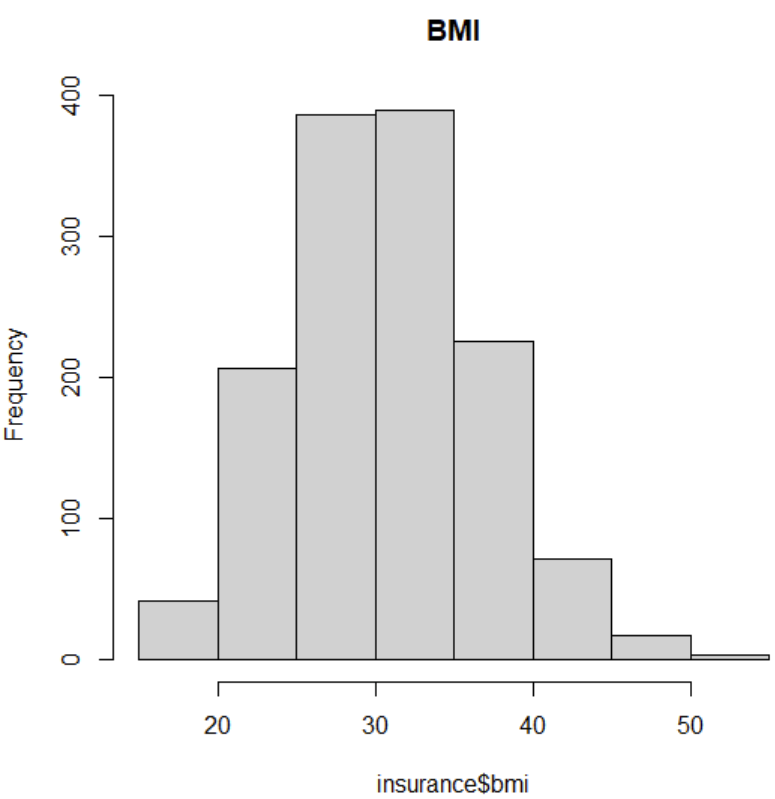
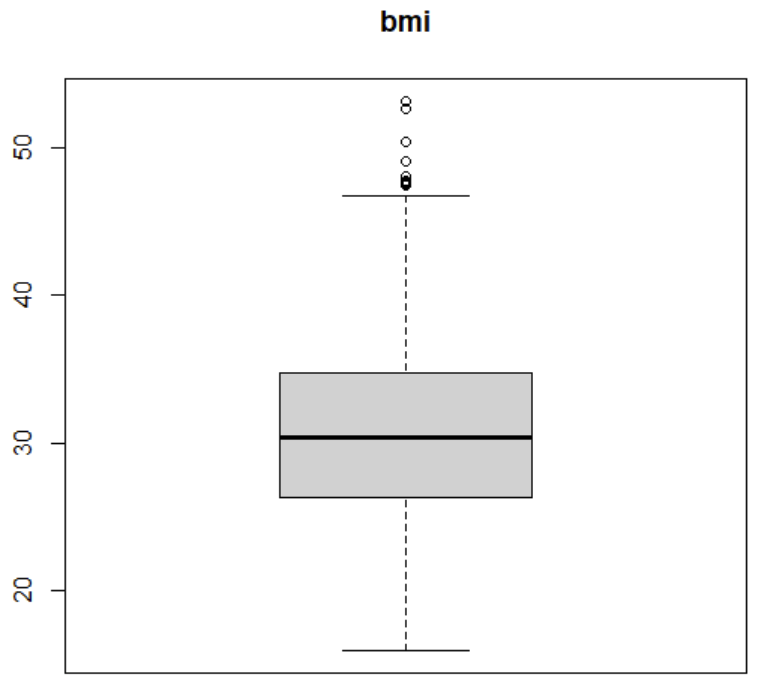


Figure B: a box plot and histogram of BMI. The histogram on the right reveals a near normal distribution, while the box plot on the left shows a handle of high outliers past BMI = ~45

Children describes the number of children/dependents that are also covered by the primary beneficiary's health insurance plan. The mean number of children is 1.095, with a minimum of 0 and a maximum of 5. This distribution is right-skewed, with 42.9% of policyholders having no children, 24.2% having one, 17.9% having two, 11.7% having three, 1.87% having four, and 1.34% having five.

Smoking habits records yes (coded 1) or no (coded 0) response to whether the policyholder smokes cigarettes. 20.48% of policyholders smoke and 79.52% do not.

The region describes the place of residence for the policyholder: this is divided into the southwest (coded 1), southeast (coded 2), northwest (coded 3), and northeast (coded 4) regions of America. The distribution of policyholders' residence region is 24.290% southwest, 27.205% southeast, 24.290% northwest, and 24.215% northeast.

Health insurance charges describe the total dollar amount of medical bills charged to the policyholder's health insurance plan in a year. The mean insurance charges are $13270 with a minimum of $1122 and a maximum of $63770. **Figure C** reveals a prominent right-skew with many high outliers.

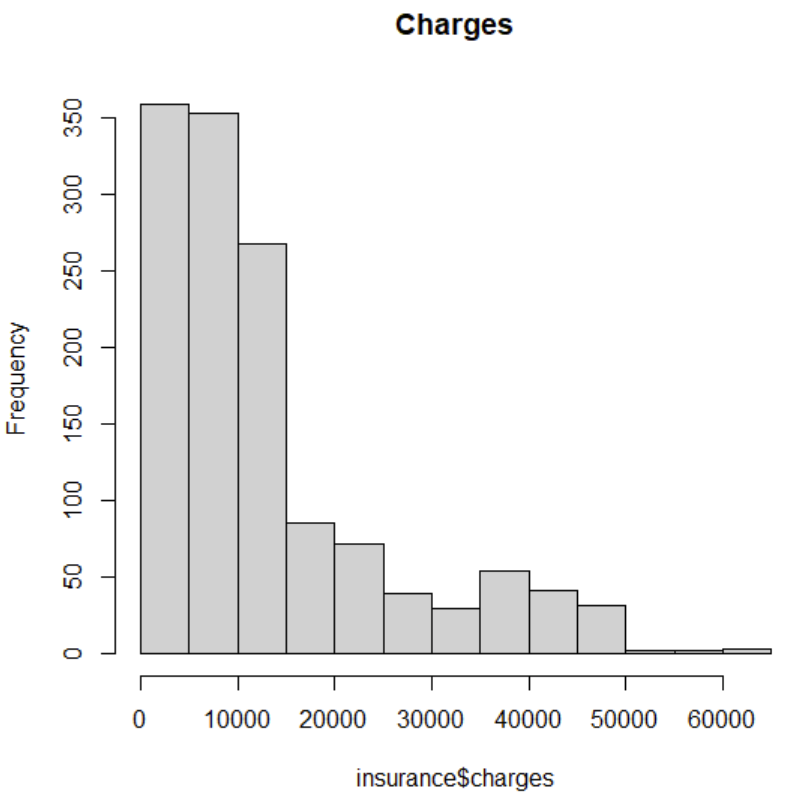
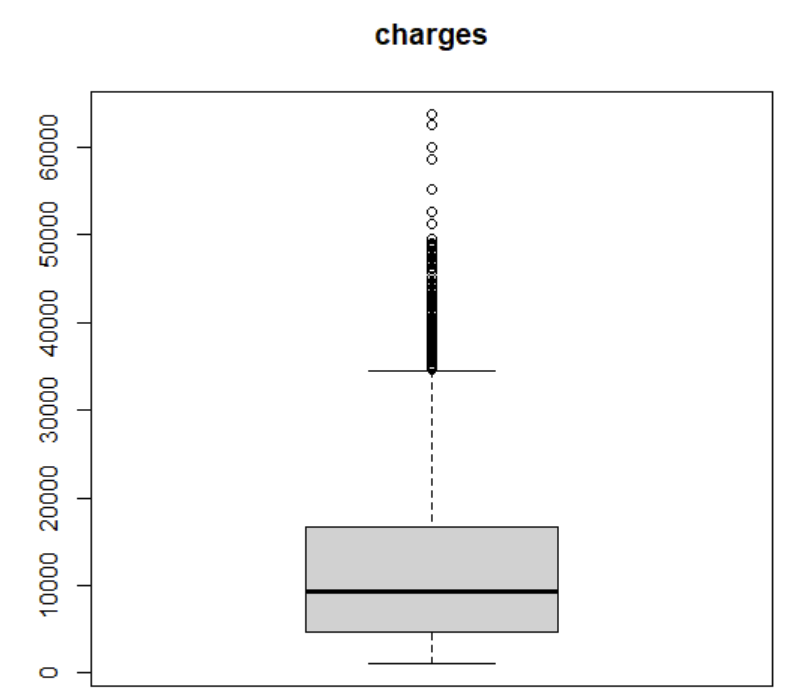


Figure C: a box plot and histogram of insurance charges. The histogram on the right reveals a right-skew distribution, while the box plot on the left confirms that the dataset contains many high outliers

The most interesting bivariate relationships are age-charge and BMI-charge. Age vs charge exhibits a strong, positive linear relationship with three main clusters. R = 0.299 for the relationship, although individual clusters appear to have a stronger correlation than indicated. BMI vs charge exhibits a weak, positive linear relationship (r = 0.198) with a gap in the center of the plot. Both predictors appear powerful, but we could not explain these unexpected features.

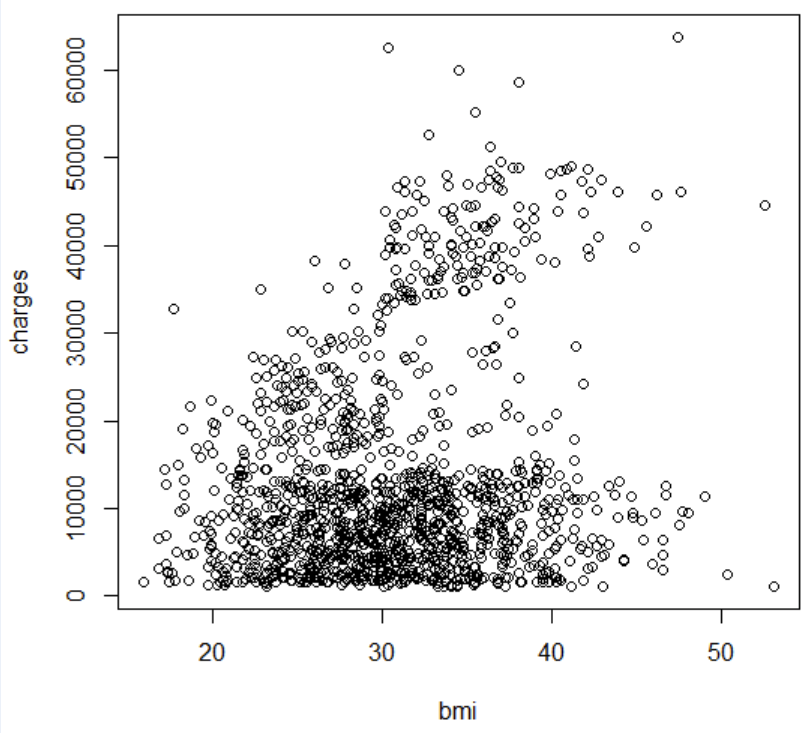
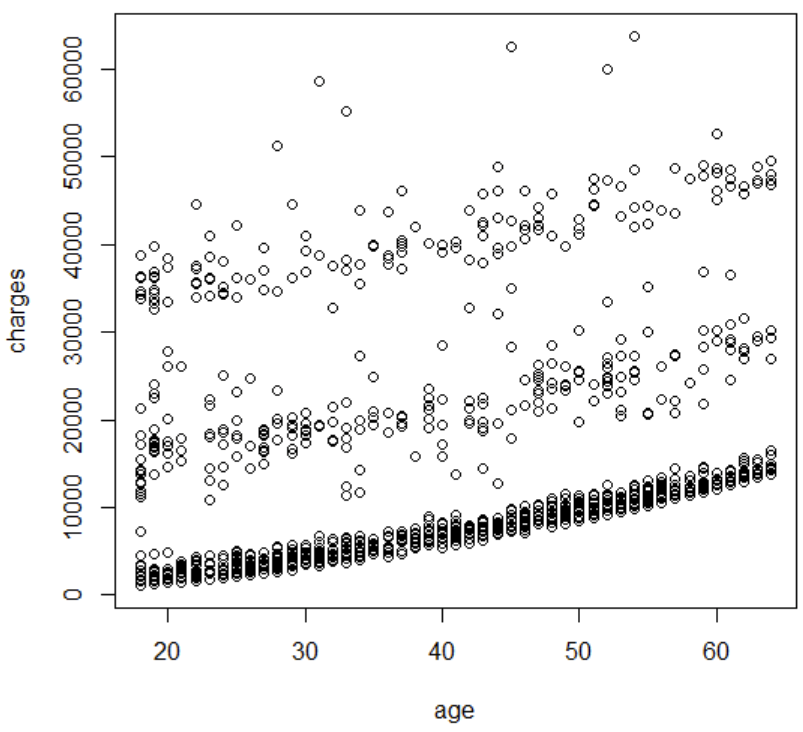


Figure D: (left) age vs charge plot, showing three distinct clusters (right) age vs bmi plot, showing a gap in the center

As health insurance charges are typically determined by the predicted health-related costs of an individual or their family, we hypothesize that health indicators will be the most influential predictors. That would include age, BMI, smoking habits, and sex in our model. We also expect children to be influential, as covering more dependents under an insurance plan increases the likelihood of a hospital visit. The only variable we do not expect to be significant is region: this does not directly relate to a policyholder’s health or family, so we expect it to have the least predictive power.

We also hypothesize that our final model will contain a lot of unexplained variation. That is because of the outliers in charge and the general unpredictability of life.

**Model Dataset**

Due to the large number of outliers in the insurance charges, we decided to remove those observations and fit a model using only the observations with insurance charges less than $20,000 (for more information on why we removed outliers, see *Additional Work*). That reduced our sample size to 1065 policyholders. Removing the outlying charges reduced the mean to $7,960 and the variance to $22,247,101. The distribution of insurance charges is still right-skewed, though to a lesser degree. We will explore transforming this variable to make the distribution more normal should the final model fail to meet the regression assumptions.

Other notable differences in the variable statistics include a decrease in smoker frequency. While 20.48% of policyholders in the full dataset smoked, that frequency was reduced to 5.82% once we removed outlying charges. Summary statistics for other variables remained similar.

To visualize the effect of removing observations with outlying charges on the predictive power of the remaining variables, we analyzed the correlation matrix of both the full and reduced datasets. Notable decreases in correlation from the full dataset to the reduced dataset include BMI-charge (r = 0.198 to r = -0.013) and smoker-charge (r = 0.787 to r = 0.499). Notable increases in correlation from the full dataset to the reduced dataset include age-charge (r = 0.299 to r = 0.602), age-BMI (r = 0.109 to r = 0.145), age-smoker (r = -0.025 to r = -0.190), and BMI-smoker (r = 0.004 to r = -0.234). The decreased correlation between predictor and response variables and the increased correlation between predictor variables indicates that multicollinearity may be a more significant issue in the reduced dataset. This dataset may be less diverse, obscuring the true predictive power of some variables and causing some variables to explain much of the same variation. We will further discuss the effects of the reduced dataset and multicollinearity in *Discussion and Limitations.*

With a significant increase in the potential predictive power of age on charges, we plotted the two variables to visualize the relationship. The plot (**Figure E**) shows a strong, positive relationship between age and charges, with a handful of influential points where younger policyholders had unexpectedly high insurance charges. The plot also exhibits a curved relationship, indicating that a quadratic term may be appropriate when fitting a model.

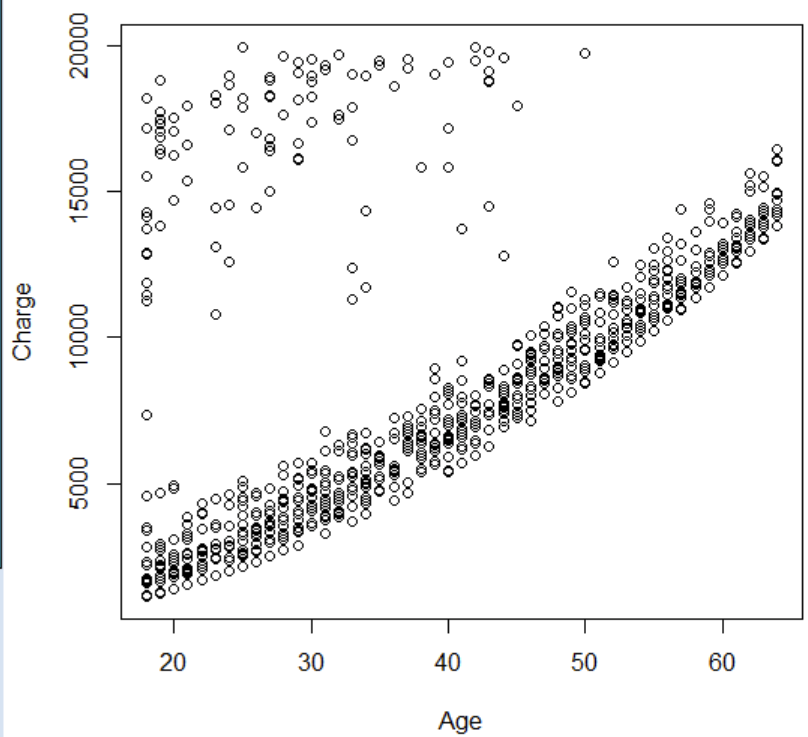


Figure E: Interaction of age and charges

**Regression Analysis**

A model of the given variables was fit before considering quadratic and interaction terms. We proceeded with a backward selection method and found that a full model with age, BMI, sex, children, smoker, and region was significant (Ho: ꞵk = 0, Ha: ꞵk ≠ 0. If p ≥ 0.05, conclude ꞵk = 0; p < 0.05, conclude ꞵk ≠ 0. All pk < 0.05 and the data supports that all predictors are significant).

*[1]* Charges = Age + BMI + Sex + Children + Smoker + Region

From here, we fit the model with polynomial and interaction terms.

As indicated from the plot of age vs charges, the relationship between the two variables appears nonlinear. A plot of age vs model *[1]* residuals (**Figure F**) confirmed that the relationship was nonlinear, as shown by the main curve near residuals = 0. As such, we fit a quadratic age term to the model. Age2 was significant (Ho: ꞵAge^2 = 0, Ha: ꞵAge^2 ≠ 0. If p ≥ 0.05, conclude ꞵAge^2 = 0; p < 0.05, conclude ꞵAge^2 ≠ 0. p = 1.58\*10-15 and the data supports Age2 is a significant predictor) in predicting charges. The marginal contribution of age2 in reducing SSE is 309863718 (SSR(Age2 | *[1]*) = SSE(Age2 + *[1]*) - SSE(*[1]*) = 5309319914 - 4999456196). This contribution reduced the previous model’s SSE by 5.836% (R2Charges Age^2 | *[1]* = SSR(Age2 | *[1]*) / SSE(*[1]*) = 309863718 / 5309319914 = 0.05836). From the p-value and marginal reduction in SSE, we concluded that Age2 should remain in the model, giving:

*[2]* Charges = Age + Age2 + BMI + Sex + Children + Smoker + Region

Once the polynomial terms were fit, we considered possible interaction terms. We explored three interactions: age-BMI, age-smoker, and BMI-smoker. We hypothesized that age, BMI, and smoker status had the greatest effect on insurance charges and reasoned that simultaneous increases in these predictors could have a more extreme impact on health. Thus, we explored each interaction pair to determine if any were significant in predicting insurance charges.

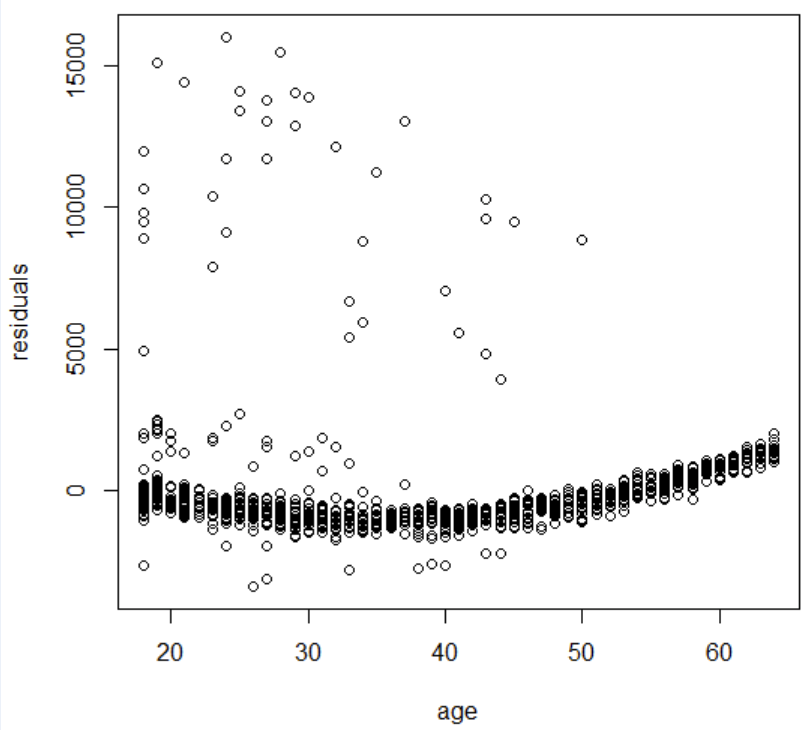


Figure F: interaction between age and residuals

We plotted each interaction term against model *[2]* residuals to determine if any should be included in the model. If any plot showed a pattern, we concluded that the interaction term accounted for some variability not already explained by the model and could be appropriately added. **Figure G** shows that the plots for age-BMI and age-smoker exhibited random scatter around 0 (except for outliers). The BMI-smoker plot exhibits a linear pattern when the policyholder is a smoker. Because the impact of BMI on charges could depend on whether the observed policyholder is a smoker, we decided to include this interaction in the model to analyze its effect.

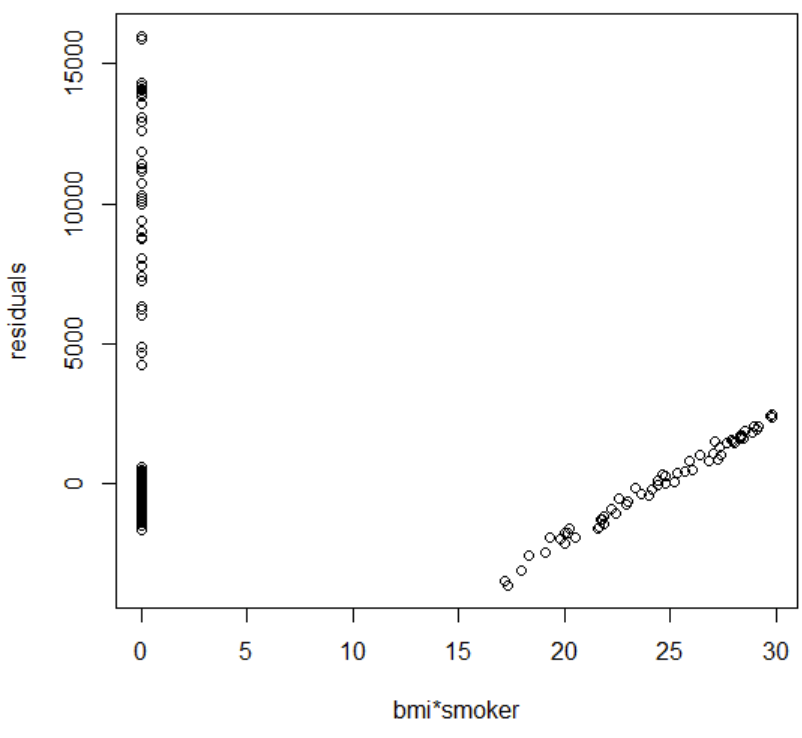
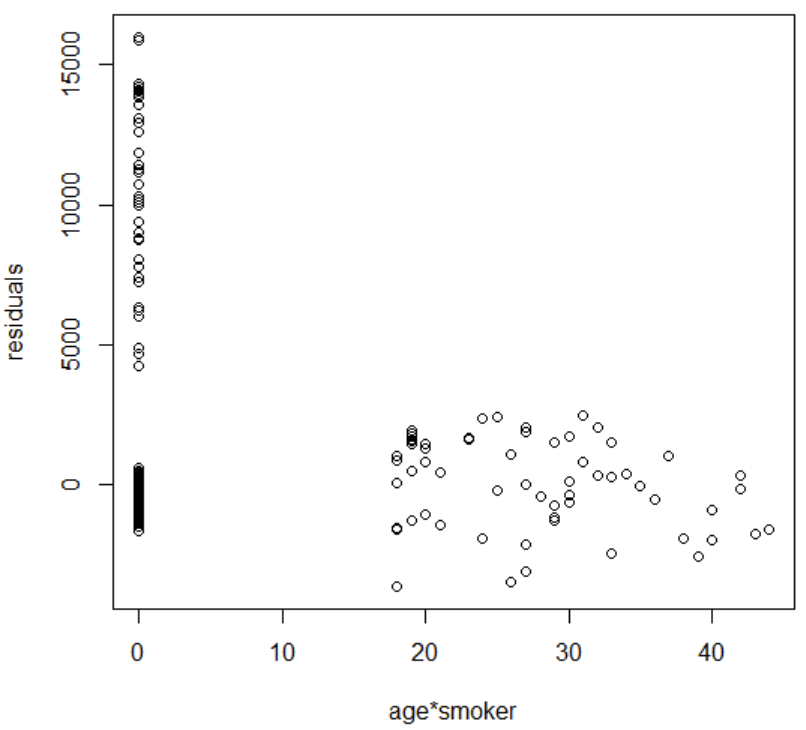
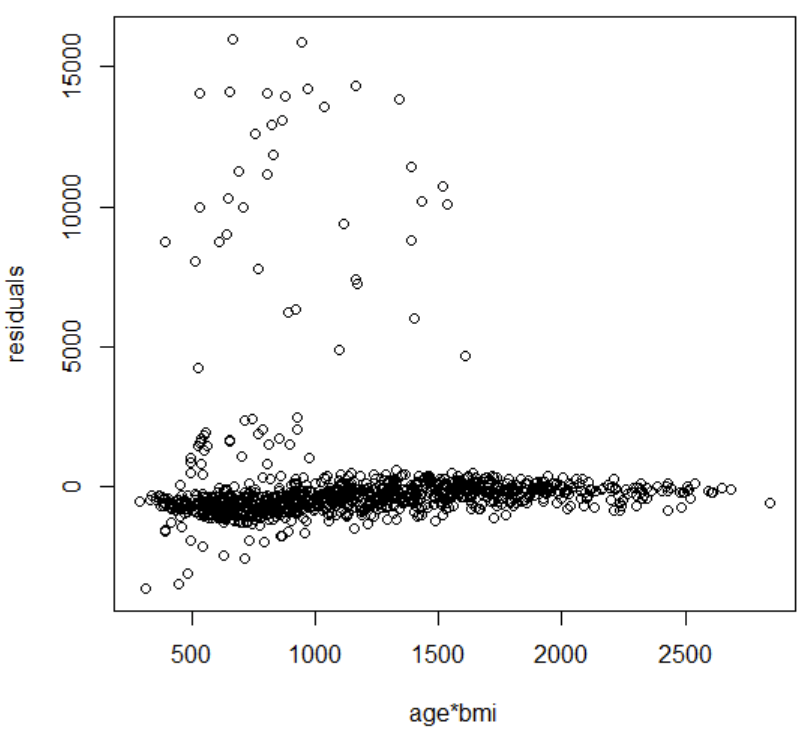


Figure G: interaction vs residuals plots for age-BMI, age-smoker, and BMI-smoker

BMI-smoker was significant (t\* = 5.923, p = 4.27\*10-9) in predicting charges. The marginal contribution of BMI-smoker in reducing SSE is 160764023 (SSR(BMI-Smoker | *[2]*) = SSE(BMI-Smoker + *[2]*) - SSE(*[2]*) = 4999456196 - 4838692173). This contribution reduced the previous models' SSE by 3.216% (R2Charges BMI-Smoker | *[2]* = SSR(BMI-Smoker | *[2]*) / SSE(*[2]*) = 160764023 / 4999456196 = 0.03216). From the p-value and marginal reduction in SSE, we concluded that the BMI-Smoker interaction should remain in the model. This gave us the final model:

*[3]* Charges = Age + Age2 + BMI + Sex + Children + Smoker + Region + BMI\*Smoker

Where age, BMI, and smoker variables were centered to avoid multicollinearity. Fitting the data to the model in terms of the original variables gives us the coefficients:

*[4]* Predicted Charges = 513.94 - 15.03(Age) + 3.26(Age2) + 22.09(BMI) - 527.07(Sex) + 645.06(Children) + 1559.04(Smoker) + 288.76(Region) + 463.40(BMI\*Smoker)

The residuals vs fitted plot (**Figure H**) revealed a cluster (n = 38) of outlying residuals for unexpectedly high charges. These outliers made up a relatively small portion of the data (3.57%), but we proceeded cautiously. The plot shows the remaining residuals randomly scattered around 0, indicating that linear regression is appropriate. Most residuals also form a horizontal band, but because of the outliers, we cannot say the error variance is constant.

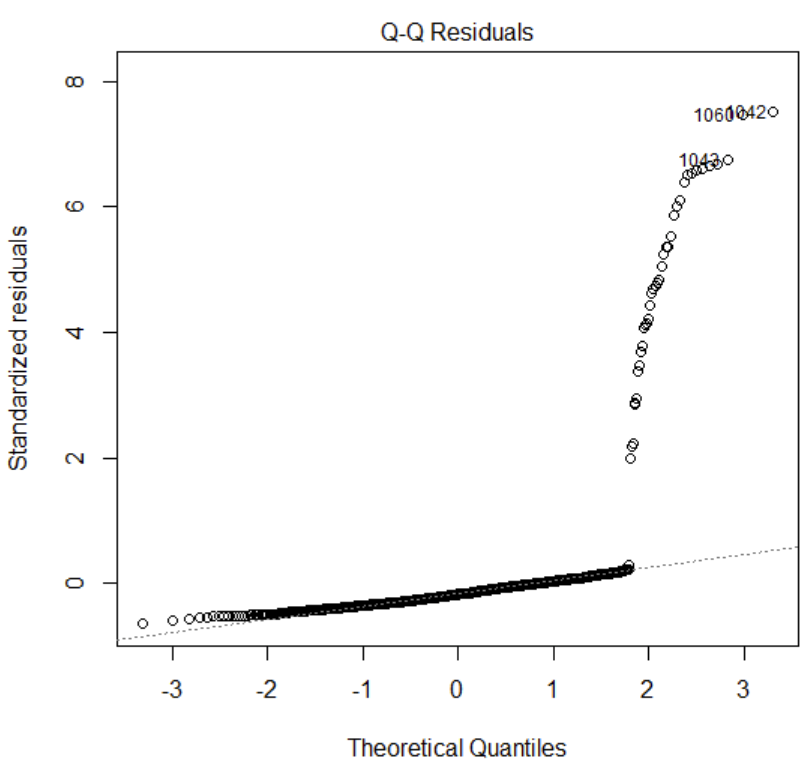
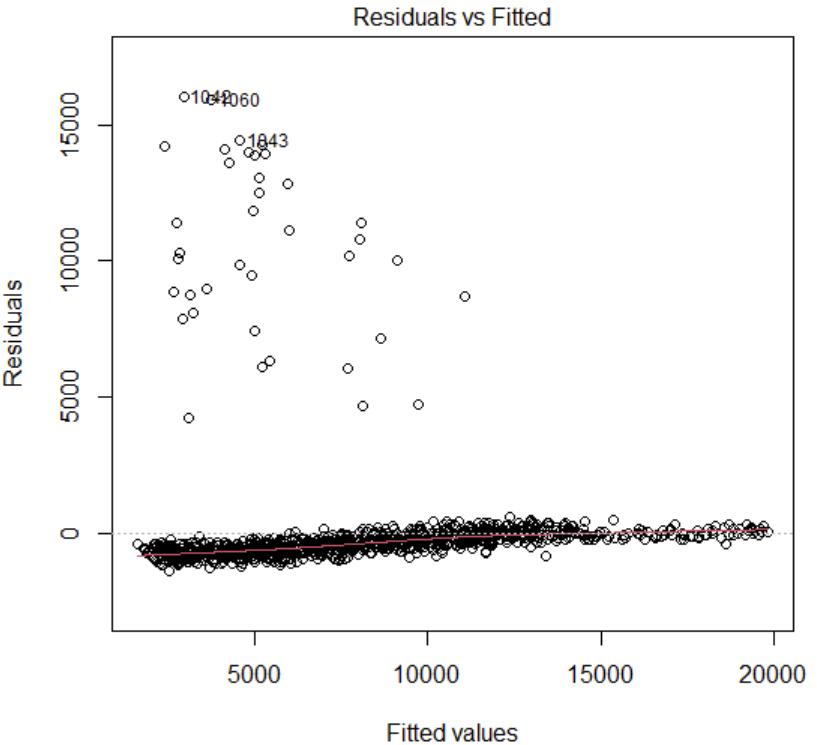


Figure H: residuals vs fitted and Q-Q residuals plots for model [4]

The QQ plot reveals that residuals skew the distribution to the right, failing the normality assumption. To fix this and the heteroscedasticity, we attempted to transform the charges to be more normal. A Box-Cox transformation calculated λ = 0.25, so we fit a new model with this transformation applied to charges. However, this did not fix the outliers or skew and reintroduced nonlinearity that was previously fixed with the age2 term (**Figure I**). We decided that–while both models fail assumptions due to outliers–*[4]* appears to fit better if we ignore the outliers for both models. We ultimately chose model *[4]*.

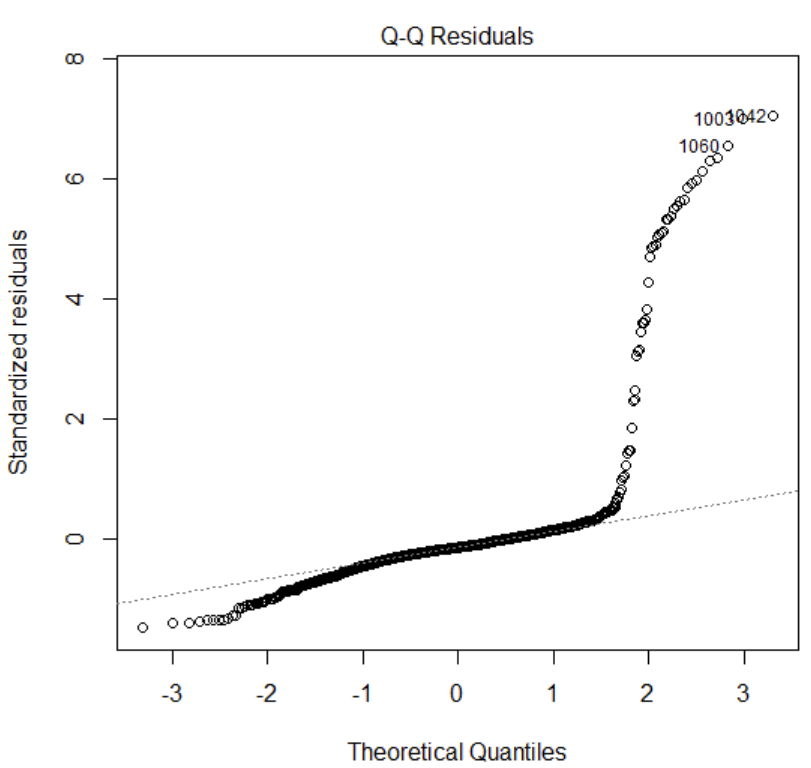
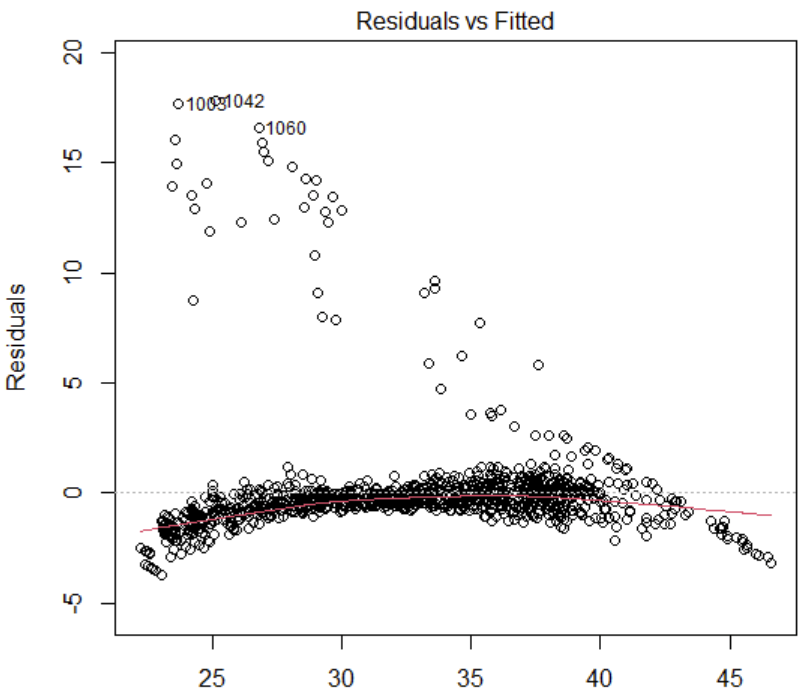


Figure I: residuals vs fitted and Q-Q residuals plots for the transformed model *[4]*

With our model chosen, we can interpret each predictor's effect on insurance charges. By taking the partial derivative of *[4]* with respect to age, we find that the mean response in charges changes by -$15.03 + $6.52(Age) per unit change in age, given that all other variables remain constant. It is interesting to note that age will initially cause the expected charges to decrease; however, because this only occurs when age is less than two and individuals must be 18 to purchase health insurance, this feature is irrelevant. A unit increase in BMI, given that all other variables remain constant, increases the expected charges by $22.09 if the policyholder does not smoke and by $485.49 (by taking the partial derivative with respect to BMI) if the policyholder does smoke. Similarly, should a policyholder begin to smoke, the mean response function for charges increases by $1559.04 + $463.4(BMI). To understand the nature of the interaction term, we plotted Charges = 513.94 + 22.09(BMI) + 1559.04(Smoker) + 463.4(BMI\*Smoker) for both smoking habits (Smoker = 1 and Smoker = 0). These functions intersected at -3.364 BMI; for larger (relevant) BMI, the expected charges for smokers were higher. The mean response function is $527.07 lower for males than for females. Similarly, the function increases by $288.76 should a policyholder move from the southwest to southeast, the southeast to northwest, or the northwest to northeast. For each additional child covered by the policyholder's insurance plan, the expected charges increase by $645.06, given that other variables remain constant.

**Discussion and Limitations**

In general, the model fits the data well. The R2 = 0.7956 indicates that 79.56% of the variation in insurance charges can be explained by model *[4]*. An F-test for model fit (Ho: there is no regression relation between charges and the predictors in *[4]*, Ha: there is a regression relation between charges and the predictors in *[4]*, conclude Ho if p ≥ 0.01 and conclude Ha is p < 0.01) calculates that *[4]* is useful in predicting charges (F\* = 513.7, p < 2.2\*10-16. At the 0.01 level, reject Ho; the data supports that there is a regression relationship). We also cross-validated our model with the K-fold method (K = 5): the resulting criterion = 4592724 compared to MSE = 4583881 from the model *[4]* output. Because these values are very close, we conclude that *[4]* performs and generalizes well when given new data. However, this model is not without its flaws.

Although we analyzed interaction terms, added explanatory variables, removed outliers, and tested a transformed response variable, we were still not able to get a fully randomly scattered residual plot or assume normality using a QQ-plot. As such, this dataset is not suited for linear modeling; nonlinear methods, like exponential regression, may be appropriate.

While removing outlying insurance charges was appropriate to attempt to fix the regression assumptions, it did not fix the error associated with estimating high charges. For example, in the tables below, we have compared a sample of data points below $20,000 in insurance charges and a sample of data points above $20,000 in insurance charges with MSE values of 316,547.5 and 167,381,724, respectively, and the same number of data points in each sample. The MSE value for the insurance charges above $20,000 was much greater, indicating significantly more variation of the data points from the predicted data points.

**Table 1:** Sample of policyholders with charges below $20,000

| age | sex | bmi | children | smoker | region | charge | fitted | error |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 46 | 0 | 32.3 | 2 | 0 | 4 | 9411.01 | 9879.39 | -468.38 |
| 26 | 0 | 34.2 | 2 | 0 | 1 | 3987.93 | 4661.28 | -673.35 |
| 56 | 1 | 33.725 | 0 | 0 | 3 | 10976.25 | 10979.82 | -3.57 |
| 34 | 0 | 27.5 | 1 | 0 | 1 | 5003.85 | 5312.78 | -308.59 |
| 23 | 1 | 27.36 | 1 | 0 | 3 | 2789.06 | 3481.44 | -692.39 |
| 18 | 0 | 31.13 | 0 | 0 | 2 | 1621.88 | 2564.82 | -942.94 |
| 54 | 1 | 24.035 | 0 | 0 | 4 | 10422.92 | 10367.38 | 55.53 |
| 62 | 1 | 39.93 | 0 | 0 | 2 | 12982.88 | 13046.02 | -63.15 |
| 29 | 1 | 31.73 | 2 | 0 | 3 | 4433.39 | 5149.98 | -716.59 |
| 21 | 1 | 23.75 | 2 | 0 | 3 | 3077.10 | 3789.94 | -712.84 |

MSE = 316547.5

**Table 2:** Sample of policyholders with charges above $20,000

| age | sex | bmi | children | smoker | region | charge | fitted | error |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 61 | 0 | 36.385 | 1 | 1 | 4 | 48517.56 | 32751.26 | 15766.30 |
| 48 | 1 | 24.42 | 0 | 1 | 2 | 21223.68 | 20768.70 | 454.98 |
| 52 | 0 | 25.3 | 2 | 1 | 2 | 24667.42 | 24257.00 | 410.42 |
| 42 | 1 | 26.07 | 1 | 1 | 2 | 38245.59 | 20544.59 | 17701.00 |
| 48 | 0 | 33.11 | 0 | 1 | 2 | 40974.16 | 25514.67 | 15459.49 |
| 47 | 1 | 28.215 | 3 | 1 | 3 | 24915.22 | 24540.40 | 374.82 |
| 18 | 0 | 36.85 | 0 | 1 | 2 | 36149.48 | 21326.51 | 14822.98 |
| 54 | 1 | 25.46 | 1 | 0 | 4 | 25517.11 | 11043.92 | 14473.19 |
| 18 | 1 | 38.17 | 0 | 1 | 2 | 36307.80 | 21440.28 | 14867.52 |
| 30 | 0 | 39.05 | 3 | 1 | 2 | 40932.43 | 26027.16 | 14905.27 |

MSE = 167381724

Although the outlier removal is appropriate, it does weaken the predictive power of the model in the case of higher insurance charges. To use the model for a higher insurance charge, in this case, you would have to rely on extrapolation. It also decreased the diversity of the dataset, especially in the case of the smokers variable where many smokers were classified as outliers.

For instance, a 95% prediction interval often fails to capture the true insurance charges for a policyholder if their observed charges are over $20000. This may occur if the observed charge is much greater than $20000. A 95% PI for the next policyholder with identical statistics to the first observation in Table 2 only expects the charges to be between $28132 and $37398; this policyholder had an actual charge of $48517. The model also cannot predict high insurance charges for non-smokers. A 95% PI for the next policyholder with identical statistics to the eighth observation in Table 2 only expects the charges to be between $6842 and $15266; this policyholder had an actual charge of $25517. The model also underestimates high charges if the age is comparatively low. A 95% PI for the next policyholder with identical statistics to the seventh observation in Table 2 only expects the charges to be between $16698 and $25959; this policyholder had an actual charge of $36149

If we were to create a new model for this data from scratch, we would try different methods to try to include the outlying data in the upper range of insurance charges, so that we would be able to use the model to predict those higher insurance charges more accurately. To do this, we may have done more meta-analyses to understand the reasons why the charges were so high for certain hospital visits. If we were to design the experiment over, we may want to test how insurance charges change over time for each individual in relation to lifestyle changes (e.g., aging, childbirth, moving to a new region, etc.).

**Conclusion**

Health insurance charges can be somewhat predicted by the variables age, BMI, number of children, smoking habits, and region, but there is evidence to suggest that there is difficulty fitting a *linear* model using these variables. Although we analyzed interaction terms, added explanatory variables, removed outliers, and tested a transformed response variable, we were still not able to get a fully randomly scattered residual plot or assume normality using a QQ-plot. The final model that fit the most after alteration [4] was Predicted Charges = 513.94 - 15.03(Age) + 3.26(Age2) + 22.09(BMI) - 527.07(Sex) + 645.06(Children) + 1559.04(Smoker) + 288.76(Region) + 463.40(BMI\*Smoker). This accounted for 79.56% of the variation in insurance charges.

**Additional Work**

Originally, we tried fitting a model to the full dataset. A backward selection method found that sex was the only nonsignificant predictor (Ho: ꞵk = 0, Ha: ꞵk ≠ 0. If p ≥ 0.05, conclude ꞵk = 0; p < 0.05, conclude ꞵk ≠ 0. psex = 0.694 and the data supports that sex is not a significant predictor. Other pk < 0.05 and the data supports that all other predictors are significant). That gives the model:

*[5]* Predicted Charges = -13280.65 + 257.41(Age) + 322.04(BMI) + 23808.21(Smoker) + 353.45(Region) + 478.44(Children)

An analysis of the fitted vs residuals plot reveals outliers and a fanning that signifies nonconstant error variance. The QQ plot is nonlinear, indicating that the residuals are not normally distributed. To attempt to fix these assumptions, we log-transformed charges to make the sample distribution more normal. However, this did not fix any of the assumptions. The fitted vs residuals plot curves (indicating nonlinearity) and funnels (indicating nonconstant error variance). The QQ plot remains nonlinear, indicating that the residuals are still non-normally distributed.

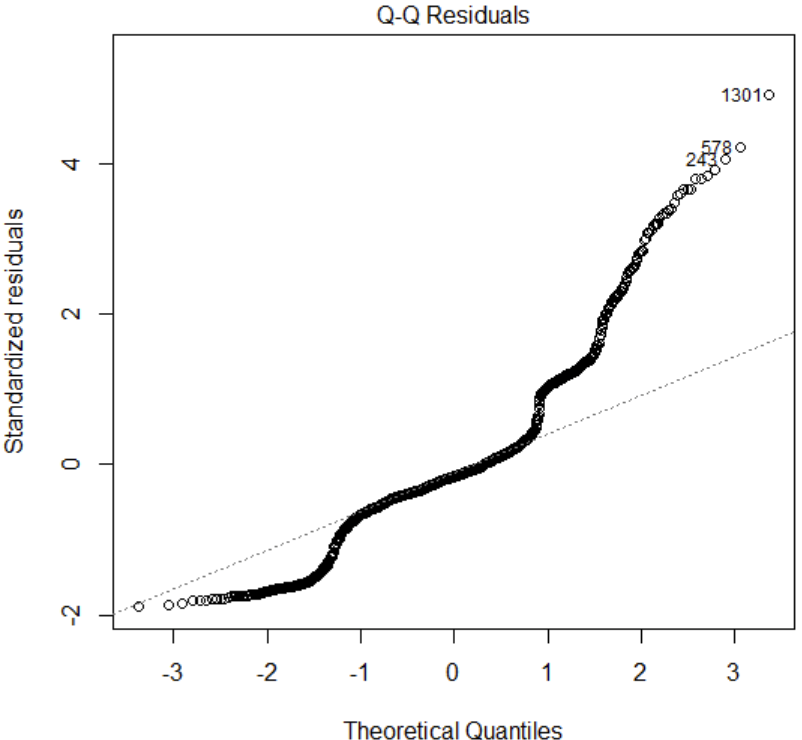
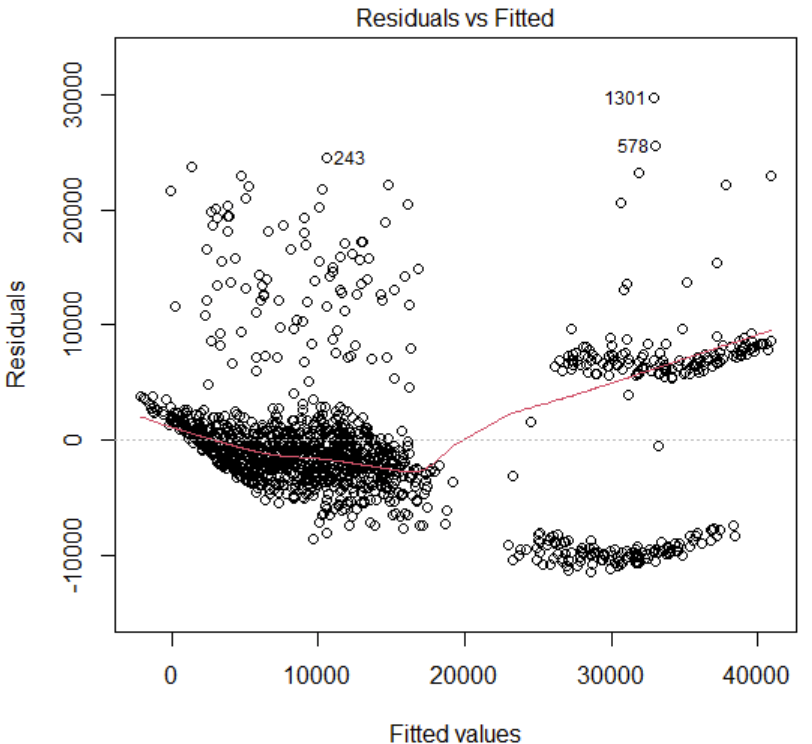


Figure J: (left) the residual vs fitted plot for *[5]*, presents outliers and heteroskedasticity (right) the QQ plot for *[5],* presents non-normal errors

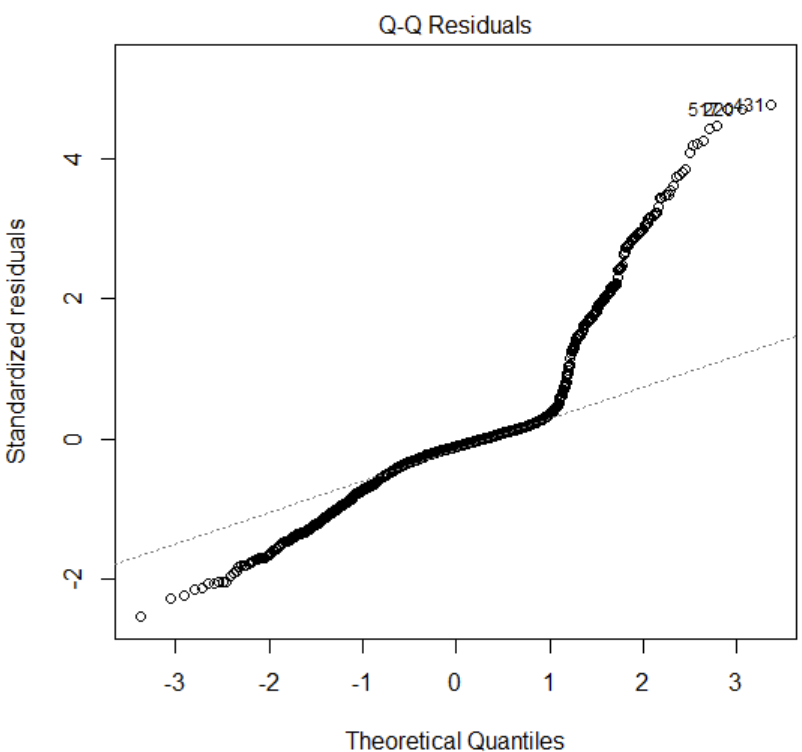
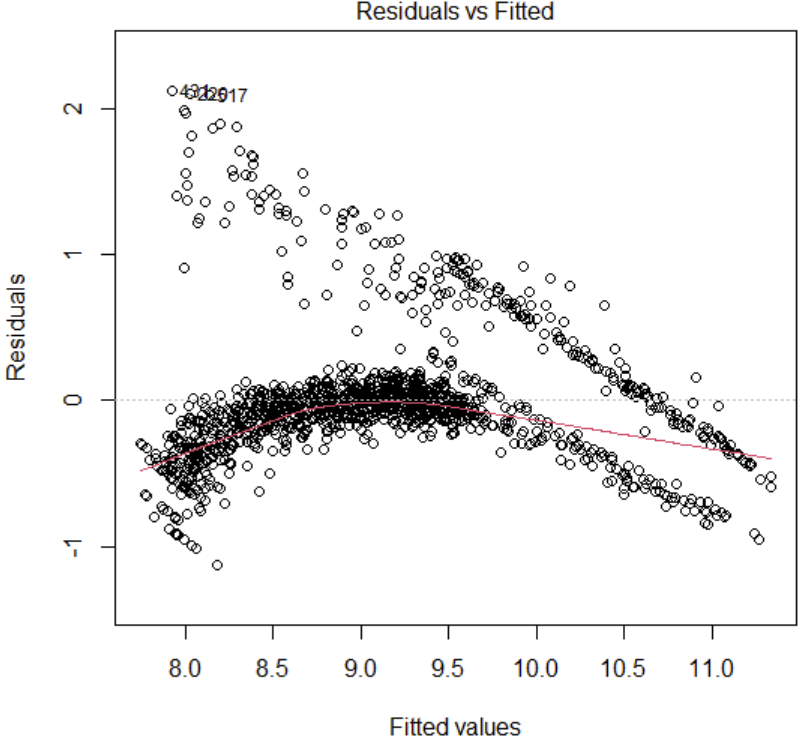


Figure K: (left) the residual vs fitted plot for log transformed *[5]*, presents nonlinearity and heteroskedasticity (right) the QQ plot for log transformed *[5],* presents non-normal errors

Because transforming that data did not work, and the data presented so many outliers, we decided to remove high outliers to try and fit the data to a better model.

**Appendix: R Code**

// insurance O is a dataset that removes the outlying charges. Charges > $20000 are

// removed

insuranceO<-read.csv(“insuranceO.csv”)

summary(insuranceO)

cor(insuranceO)

plot(insuranceO)

// plot of age vs charges appears to curve slightly, quadratic term may be helpful

// fit a full model of the data

modelOFull<-lm(charges~age+bmi+sex+children+smoker+region, data=insuranceO)

summary(modelOFull)

// all predictors are significant

plot(insuranceO$age, modelFull$residuals)

// the plot of age vs residuals appears to curve, indicating that a quadratic term would be helpful

insuranceO$ageC<-(insuranceO$age-mean(insuranceO$age))

modelOFullQ<-lm(charges~ageC+I(ageC^2)+bmi+sex+children+smoker+region, data= insuranceO)

summary(modelOFullQ)

// adding age2 to the model, significant

anova(modelOFull)

anova(modelOFullQ)

// SSR(Age2 | *[1]*) = SSE(Age2 + *[1]*) - SSE(*[1]*) = 5309319914 - 4999456196 = 309863718

// R2Charges Age^2 | *[1]* = SSR(Age2 | *[1]*) / SSE(*[1]*) = 309863718 / 5309319914 = 0.05836

plot(I(insuranceO$bmi\*insuranceO$smoker), modelOFullQ$residuals)

plot(I(insuranceO$age\*insuranceO$smoker), modelOFullQ$residuals)

plot(I(insuranceO$age\*insuranceO$bmi), modelOFullQ$residuals)

// the interaction plot between bmi and smoker presents a linear pattern, indicating that a

// bmi\*smoker term should be added to the model

insuranceO$bmiC<-(insuranceO$bmiC-mean(insuranceO$bmiC))

insuranceO$smokerC<-(insuranceO$smokerC-mean(insuranceO$smokerC))

modelO<-lm(charges~ageC+I(ageC^2)+bmiC+smokerC+children+sex+I(bmiC\*smokerC), data=insuranceO)

// adding bmi\*smoker to the model, significant – this is the final model

summary(modelO)

anova(modelO)

// SSR(BMI-Smoker | *[2]*) = SSE(BMI-Smoker + *[2]*) - SSE(*[2]*) = 4999456196 - 4838692173 =

// 160764023

// R2Charges BMI-Smoker | *[2]* = SSR(BMI-Smoker | *[2]*) / SSE(*[2]*) = 160764023 / 4999456196 = 0.03216

plot(modelO)

//The residuals vs fitted plot contains several high outliers, the error distribution is non-normal

library(MASS)

boxcox(modelO)

// 𝜆 = 0.25

insuranceO$chargesT<-((insuranceO$charges^0.25 - 1)/0.25)

modelOT<-lm(chargesT~ageC+I(ageC^2)+bmi+sex+children+smoker+region+I(smoker\*bmi),data=insuranceO)

plot(modelOT)

// the outlying residuals are not fixed, residuals vs fitted appear curved, and errors are still

// non-normal. This model was abandoned for modelO

install.packages("cv")

library(cv)

cv(modelO,k=5,seed=1234)

// K-fold cross-validation for the model

modelO2<-lm(charges~age+I(age^2)+bmi+sex+children+smoker+region+I(smoker\*bmi),data=insuranceO)

summary(modelO)

// the final model in terms of the original predictor variables

library(dpylr)

subLow<-insurance[insurance$charges < 20000,]

sampleLow<-sample\_n(subLow,10)

subHigh<-insurance[insurance$charges > 20000,]

sampleHigh<-sample\_n(subHigh,10)

sampleLow$fitted<-(513.94-(15.03\*sampleLow$age)+(3.26\*(sampleLow$age^2))+(22.09\*sampleLow$bmi)-(527.07\*sampleLow$sex)+(645.06\*sampleLow$children)+(1559.04\*sampleLow$smoker)+(288.76\*sampleLow$region)+(463.4\*sampleLow$bmi\*sampleLow$smoker))

sampleHigh$fitted<-(513.94-(15.03\*sampleHigh$age)+(3.26\*(sampleHigh$age^2))+(22.09\*sampleHigh$bmi)-(527.07\*sampleHigh$sex)+(645.06\*sampleHigh$children)+(1559.04\*sampleHigh$smoker)+(288.76\*sampleHigh$region)+(463.4\*sampleHigh$bmi\*sampleHigh$smoker))

// fitted with coefficients in terms of original variables

sampleLow$error<-(sampleLow$charges-sampleLow$fitted)

sampleHigh$error<-sampleHigh$charges-sampleHigh$fitted)

// calculating the errors for a subsample of the data – both data used to fit the model and outliers

mseLow<-(sum(sampleLow$error^2)/10)

mseHigh<-(sum(sampleHigh$error^2)/10)

//The MSE for observations with low charges (316547.5) is much lower than that of high,

// outlying charges (167381724). One limitation of our model is that is cannot accurately predict

// high insurance charges

predict(modelO2,newdata=data.frame(age=54,sex=1,bmi=25.46,children=1,smoker=0,region=4),interval="prediction",level=0.95)

predict(modelO2,newdata=data.frame(age=61,sex=0,bmi=36.385,children=1,smoker=1,region=4),interval="prediction",level=0.95)

predict(modelO2,newdata=data.frame(age=18,sex=0,bmi=36.85,children=0,smoker=1,region=2),interval="prediction",level=0.95)

//The model fails to accurately predict high charges, especially if the policyholder does not

// smoke (age 54), the charge is particularly high (age 61), or the person is young compared to

// the charge (age 18)

// Additional Work – model with the full dataset

insurance<-read.csv(“insurance.csv”)

summary(insurance)

plot(insurance)

cor(insurance)

model<-lm(charges~age+bmi+sex+children+smoker+region, data=insurance)

summary(model)

// sex is nonsignificant, dropped

model<-lm(charges~age+bmi+children+smoker+region, data=insurance)

plot(model)

// the plot shows issues of heteroskedasticity and nonnormality, transform with log(charges)

modelT<-lm(I(log(charges))~age+bmi+children+smoker+region, data=insurance)

plot(modelT)

// the assumptions still fail, and at this point the model was abandoned for our final model